IAIC Transactions on Sustainable Digital Innovation (ITSDI)
Vol. 1 No. 1 October 2019

Optimization the Naive Bayes Classifier Method to diagnose diabetes Mellitus

Desi Susilawati[1], Dwiza Riana[2]
Infomation techlonogy, Universitas Bina Sarana Informatika[1], Computer Science, STMIK Nusa Mandiri[2]
Indonesia
e-mail:desi.dlu@bsi.ac.id, dwiza@nusamandiri.ac.id

To cite this document:
Susilawati, D. S., & Riana, D. (2019). Optimization the Naive Bayes Classifier Method to diagnose diabetes Mellitus. *IAIC Transactions on Sustainable Digital Innovation (ITSDI)*, *1*(1), 78-86. Retrieved from https://aptikom-journal.id/index.php/itsdi/article/view/21

Author Notification
12 October 2019
Final Revised
18 October 2019
Published
21 October 2019

Abstract

World Health Organization (WHO) states that Diabetes Mellitus is the world's top deadly disease. several studies in the health sector including diabetes mellitus have been carried out to detect diseases early. In this study optimization of naive bayes classifier using particle swarm optimization was applied to the data of patients with 2 classes namely positive diabetes mellitus and negative diabetes mellitus and data on patients with 3 classes, those who tested positive for diabetes mellitus type 1, diabetes mellitus type 2 and negative diabetes mellitus. After testing, the algorithm of Naive Bayes Classifier and Naive Bayes Classifier based on Particle Swarm Optimization, the results obtained are the Naive Bayes Classifier method for 2 classes and 3 classes each producing an accuracy value of 78.88% and 68.50%. but after adding Particle Swarm Optimization the value of accuracy increased respectively to 82.58% and 71, 29%. The classification results for 2 classes have an accuracy value higher than 3 classes with a difference of 11.29%

Keywords: Diabetes Mellitus, naive bayes classifier, Particle Swarm Optimization

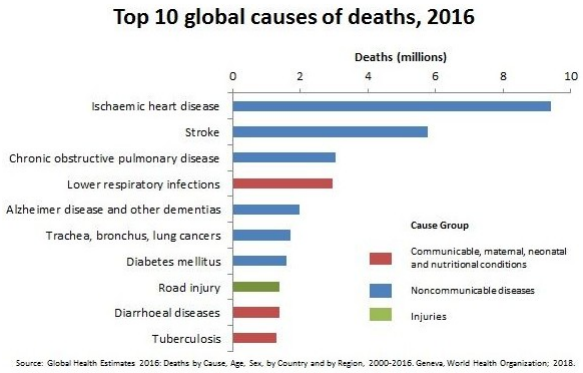
1. Introduction

An increase in sugar levels is one of the characteristics of diabetes mellitus, which leads to serious impact to the heart, blood vessels, eye, kidney, and nerve from time to time [1] diabetes is a disease that is difficult to cure.for people with diabetes, they must control blood sugar levels [2].

Diabetes is a serious disease that makes people with diabetes have to regulate blood sugar levels. one characteristic of type I diabetes mellitus is the rejection of insulin production in the body. Type II diabetes has one characteristic of insulin resistance. Type II diabetes can also cause several other serious diseases [3].

According to the updated data from WHO on 2018, from 56.9 million death in the world in 2016, more than a half of them (54%) cause by ten most disease. Heart attack ischemic and stroke is the top killer disease in the world, with totally killed 15.2 million death from combining

both diseases. This disease is still the global major cause to death in the last 15 years. However, Diabetes Mellitus is included the 10 major death cause, the disease is on the seventh that death and killed 1.6 million people, increased as it was less than a million in 2000.



Picture 1. Graphic 10 the top death cause in 2016
Source: WHO (Updated May 2018)

In order to overcome this challenge, there were plenty researches have been taken, to predict the disease accurately. However, the most accurate method has not been figured yet to predict the disease.

Naive bayes is the best and very practical method [4]. A Naive Bayes classifier is a probabilistic classifier based and can be used to classify diabetes mellitus [5]. To develop the existed research, naive bayes will be combined with Particle Swarm Optimization (PSO) for optimizing the attribute selection to enhance the accuracy of Diabetes Mellitus prediction.

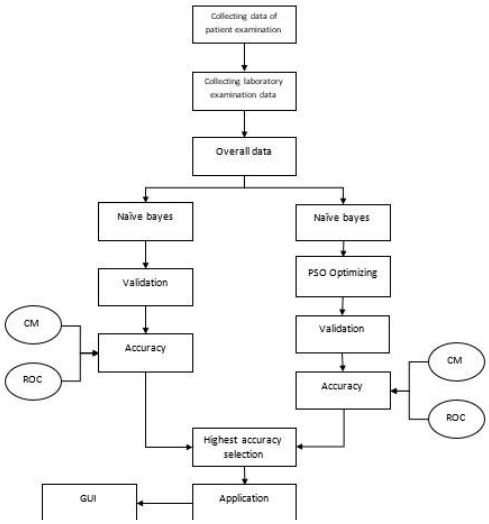
According to the description above, for reducing the weaknesses exist there will be implemented the naive bayes method combined with PSO. The optimization method will be used to enhance the Naive Bayes Classifier accuracy for predicting the Diabetes Mellitus

2. Research Method

Two Main approaches that will be used in this research are Quantitative approach and Qualitative Approach. Quantitative approach will be based on research framework, the research will explain the process on how the answer given according the research question and must be given the conceptual definition temporary. Quantitative research is typically associated with the process of enumerative induction. One of its main purposes is to discover how many and what kinds of people in the general or parent population have a particular characteristic which has been found to exist in the sample population. [6].

The Qualitative approach is different with the Quantitative approach, The Qualitative approach is scientific research that learn about condition and process. Inductive analysis research included within details and special data to find the main categories, holistic research, and symptom seek as one complex system exceeding the total of each parts. In this research, the research model is using the experiment model, where the model is involved research to some variable using specific test that controlled by the researcher. This model is testing the validity of one hypothesis with statistic and connection with the main subject of the research.

The purpose of using the model is to predict the Diabetes Mellitus according to the data set and variable that has been defined by using the Naive Bayes algorithm dan PSO, that will be used to optimize the attribute selection with the step that will be explained in research framework (picture 2) as below:



Picture 2. Research step chart

The explanation for the research step chart on picture 2 as below:

1. The research is using the patient examination result data from the Doctor and laboratory result obtained from RS Betha Medika Sukabumi. The sample of the examination result can be seen in the attachment 3rd sheet. After the Doctor examination result been obtained, the researcher collects the same patient data from the laboratory result, the sample of the laboratory result can be seen in attachment 4th.
2. After all the data had been collected, then the testing carried out by using the Naïve Bayes method and the Naïve Bayes Method with PSO optimization.
3. The result from each testing will be validate by using the same tools which is rapid miner studio 8.2
4. According to the testing evaluation, the accuracy will be count by two measurements, according to the Confusion Matrix and ROC evaluation. After the accuracy result has been achieved, each data's will be compared to get the highest score of accuracy, later will be applicated on making the Graphical User Interface (GUI).

2.1 Formula/Algorithm

The Naive Bayes classifier is a probabilistic classification method from the Bayes application theory using a strong assumption of independence among these features. [7].

$$P(X \mid Y) = \frac{P(Y \mid X)\, P(X)}{P(Y)}$$

Y = data with unknown classes
X = Y data hypothesis represents a specific class
P (X | Y) = probability of hypothesis X based on condition Y
P(X) = probability of hypothesis X
P(Y | X) = Y probability based on the conditions in hypothesis X
P(Y) = probability of Y

2.2 Literature Review

Plenty of researches to predict the Diabetes Mellitus disease, one of his studies

proposed using AdaBoost algorithm with Decision Stump as base classifier for classification. Support Vector Machine, Decision Tree and Naive Bayes are also used to classify predictions of diabetes mellitus which are also used with the AdaBoost algorithm for verification of accuracy. Accuracy obtained for AdaBoost algorithm with a decision stump is 80.72% and includes high accuracy [2].

The next research is Prediction using two types of automatic tools like ANN and the time series base system. Time series based prediction for various patients data (UCI repository) which is public data. predictions using Time series based have a difference of 1 to 5%. while predictions using ANN based prediction have a difference of 19 to 22%. then it can be concluded that the Time series based has better results compared ANN based prediction [8]. The next research is show evaluation of the results of the adaboost ensemble method, outperforms than bagging as well as standalone J48 decision tree. The adaboost ensemble method can also be used for classification of other disease data such as coronary heart disease, hypertension and dementia [9].

3. Findings

3.1 Problem

1. Result of patient examination from RS Betha Medika Sukabumi in total 156 people suffer from diabetes mellitus, to decrease the Diabetes Mellitus risk they need to do the medical checkup foreseen the disease early. The utilization classification technique with high accuracy and precise prediction able to assist the issue above. So that the diagnosis result can be faster, easier and more accurate. In this research, we are implementing the optimization algorithm naïve bayes classifier with attribute selection using PSO.
2. The data that has been used is the primary data from the patient examination result in RS Betha Medika Sukabumi, which consist of six predictor attributes and one result attribute from the data can be seen the positive status patient and negative status patient and the differentiation classification of Diabetes Mellitus type 1 and Diabetes mellitus type 2.

3.2 Research Implementation

A. Result of Naïve Bayes Classifier method

Table 1. data of patient RS Betha Medika Sukabumi in total 156 persons (2 class)

Name	Statistics	Range
Class	mode = Yes (110), least = No (46)	Yes (110), No (46)
Pregnance	mode = Medium (58), least = Low (43)	Medium (58), High (55), Low (43)
Glucose Level	mode = Medium (75), least = Low (21)	Medium (75), Low (21), High (60)
Blood Pressure	mode = Normal (84), least = High (18)	High (18), Normal (84), Normal to High (36), Normal to high (18)
Body Index Mass	mode = Severely Obese (88), least = Low (14)	Severely Obese (88), Obese (31), Normal (23), Low (14)
Diabetic History	mode = Low (90), least = High (66)	High (66), Low (90)
Age	mode = Young (96), least = Old (7)	Young (96), Medium (53), Old (7)

1. Evaluation model with Confusion Matrix Confusion matrix model will form a matrix that consist of true positive or tuple positive and true negative or tuple negative. Later, input the data exist inside the confusion matrix so we obtained the result as below:

Table 2. Confusion Matrix method Naive Bayes Classifier

accuracy: 78.88% +/- 7.54% (micro average: 78.85%)

	true Ya	true Tidak	class precision
pred. Ya	95	18	84.07%
pred. Tidak	15	28	65.12%
class recall	86.36%	60.87%	

According to the above table, the result True Positive (TP) is 95, False Negative

(FN) is 28, False Positive (FP) is 15 dan True Negative (TN) is 18, from those data we able to calculate the score of accuracy, sensitivity, specificity, PPV dan NPV. The result of processing data on below table:

Table 3. Score of Accuracy, Sensitivity, Specificity, PPV and NPV

	Value
Accuracy	0, 78846
Sensitivity	0. 8407
Specificity	0. 6511
PPV	0. 8636
NVP	0. 6086

Manually the data can be calculate using equation as below:

Accuracy $= \frac{95+28}{95+28+15+18} = 0{,}78846$

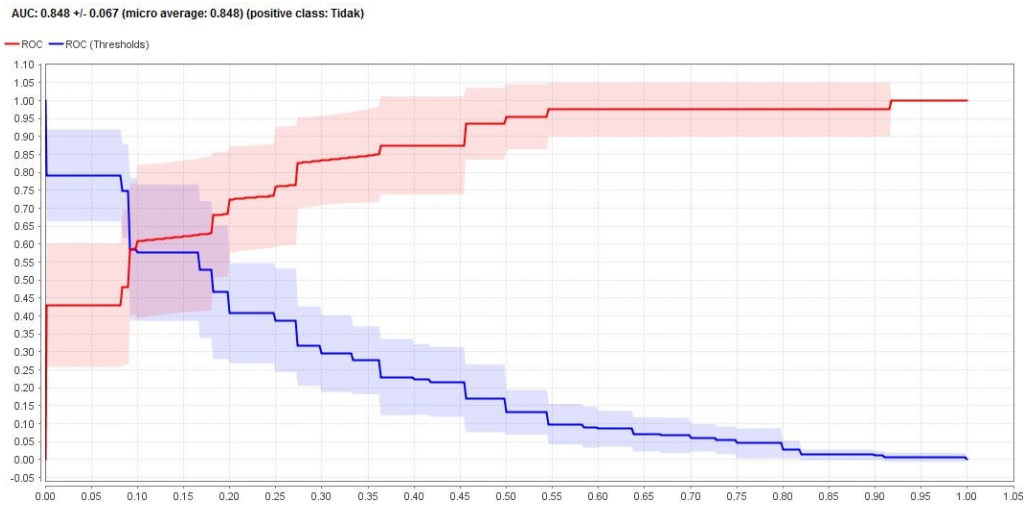
Sensitivity $= \frac{95}{95+18} = 0{,}8407$

Specificity $= \frac{28}{28+15} = 0{,}6511$

PPV $= \frac{95}{95+15} = 0{,}8636$

NPV $= \frac{28}{28+18} = 0{,}6086$

2. Evaluation using ROC Curve The test result from the data of Naïve Bayes classifier Method against the ROC scorecard can be seen in below picture:



Picture 3. AUC score naïve bayes method in ROC graphic

AUC (Area Under Curve) Score 0.848 for evaluation method of naïve bayes classifier

shown the score as fair classification.

B. Result of Naïve Bayes Classifier method Using PSO

1. Evaluation model with Confusion Matrix The result from trial that has been done for rating accuracy score and AUC Score by using the naïve bayes classifier with PSO as below:

Table 4. Confusion Matrix naïve bayes classifier method based with PSO

accuracy: 82.58% +/- 10.89% (micro average: 82.69%)

	true Ya	true Tidak	class precision
pred. Ya	97	14	87.39%
pred. Tidak	13	32	71.11%
class recall	88.18%	69.57%	

According to the table above, the result has been released as True Positive (TP) 97, False Negative (FN) 14, False Positive (FP) 13 dan True Negative (TN) 32. From those data we can be able to score the accuracy, sensitivity, specificity and NPV. the result of processed data can be seen in below table:

Table 5. Score of Accuracy, Sensitivity, Specificity, PPV and NPV

	Value
Accuracy	0, 8269
Sensitivity	0, 8738
Specificity	0, 6511
PPV	0, 8818
NPV	0, 6956

The data can be calculate manually using the equation as below:

Accuracy $= \frac{97+32}{97+32+13+14}$ = 0,8269

Sensitivity $= \frac{97}{97+14}$ = 0,8738

Specificity $= \frac{32}{32+13}$ = 0,6511

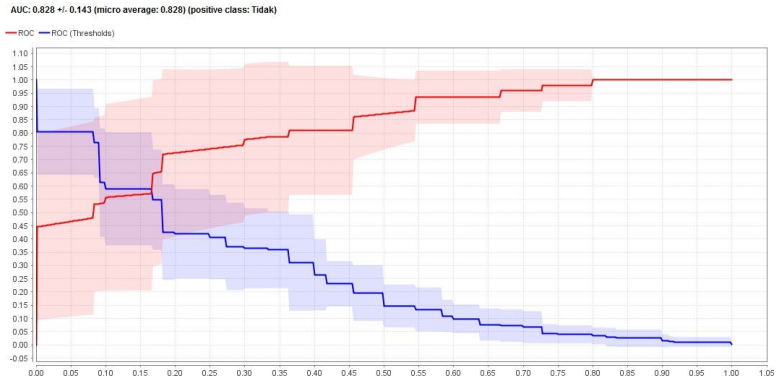
PPV $= \frac{97}{97+13}$ = 0,8818

NPV $= \frac{32}{32+14}$ = 0,6956

According to the scoresheet, can be acknowledged that the sensitivity score has the lowest score compare to other point.

2. Evaluation with ROC Curve

The testing result of naïve bayes classifier method with PSO Based against the ROC scorecard in below picture:



Picture 4. AUC Score of naïve bayes classifier with PSO based in ROC Graphic

According to the AUC Score of 0.828 as seen as above picture, so the score of accuracy has Good Classification level.

C. Evaluation analysis and Validation model

According the data result above, the evaluation either based on confusion matrix or ROC curve proven that the testing that used naïve bayes classifier with PSO based has higher

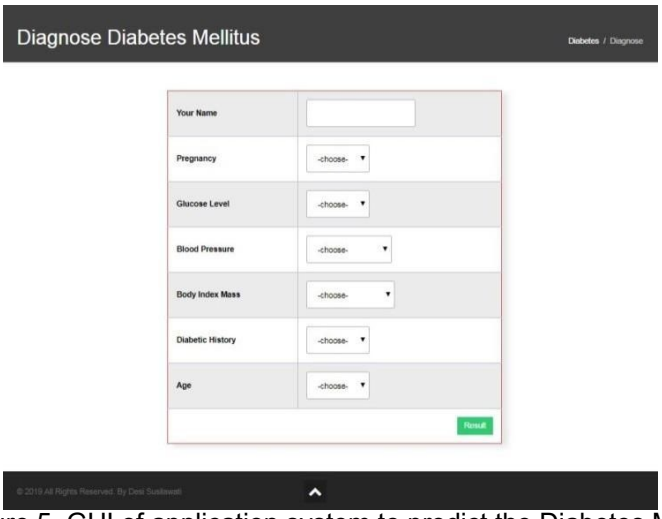
accuracy compare to only using naïve bayes classifier. The accuracy score of naïve bayes classifier algorithm method on two classes reach 78.88% with the AUC score 0.848 and the score of naïve bayes classifier with PSO Based for two classes reach 82.58% with AUC score 0.828. While the accuracy score of naïve bayes classifier algorithm method for three classes reach 68.50% and for the score of naïve bayes classifier with PSO based for three classes reach 71.29%. Based on the score above obtained the different accuracy level for two classes and three classes 11.29% can be seen in below table 6:

Table 6. Algorithm classification testing of naïve bayes classifier and naïve bayes classifier with PSO based

	2 classes	3 classes
Naive Bayes Classifier	78,88 %	68,50 %
Naive Bayes Classifier with PSO based	82,58%	71, 29%

D. The measurement result of GUI reliability test

According to the evaluation of optimizing naïve bayes classifier with PSO method has higher score result than the naïve bayes classifier method only, so the rule that has been produce by naïve bayes classifier with PSO will be selected to be the rule of producing Graphical User Interface in order to assist the doctor and medical assistant or civil society for diagnosing or predict the Diabetes mellitus. The Interface used in this research could be seen at below picture:



Picture 5. GUI of application system to predict the Diabetes Mellitus

4. Conclusion

From the above Discussion we can summarize that:

- The research with particle swarm optimization implementation for selecting the attribute data in naïve bayes classification able to enhance the accuracy to diagnose the Diabetes Mellitus.
- For selecting the attribute needed of particle swarm optimization algorithm in naïve bayes, the learning process experimented with Rapidminer. While the accuracy test is using Accuracy and ROC Curve for seeking the accuracy level of diagnostic Diabetes Mellitus prediction.
- The result score from the research on naive bayes classifier method is 71,79 % with AUC score 0,739 and for the accuracy score on PSO based naïve bayes classification is 89, 74% with AUC score 0,876. According to the above score, obtained difference 17.95% and difference in AUC 0.13.

So that can be concluded the optimizing technique by particle swarm optimization able to select the naïve bayes attribute and produce higher level of accuracy in diagnosing Diabetes Mellitus better than just individual naïve bayes method.

References

[1] WHO. (2016). *WHO*. Retrieved from diabetes programme. http://www.who.int/diabetes/en/

[2] Vijayan, V. V., & Anjali, C. (2015, December). Prediction and diagnosis of diabetes mellitus—A machine learning approach. In *2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)* (pp. 122-127). IEEE.

[3] Zaccardi, F., Webb, D. R., Yates, T., & Davies, M. J. (2016). Pathophysiology of type 1 and type 2 diabetes mellitus: a 90-year perspective. *Postgraduate medical journal*, *92*(1084), 63-69.

[4] Gao, C. Z., Cheng, Q., He, P., Susilo, W., & Li, J. (2018). Privacy-preserving Naive Bayes classifiers secure against the substitution-then-comparison attack. *Information Sciences*, *444*, 72-88.

[5] Vijayarani, S., & Dhayanand, S. (2015). Liver disease prediction using SVM and Naïve Bayes algorithms. *International Journal of Science, Engineering and Technology Research (IJSETR)*, *4*(4), 816-820.

[6] Brannen, J. (2017). Mixing Methods. Qualitative and Quantitative Research.

[7] Vembandasamy, K., Sasipriya, R., & Deepa, E. (2015). Heart diseases detection using Naive Bayes algorithm. *International Journal of Innovative Science, Engineering & Technology*, *2*(9), 441-444.

[8] Rani, S., & Kautish, S. (2018, June). Association Clustering and Time Series Based Data Mining in Continuous Data for Diabetes Prediction. In *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1209-1214). IEEE.

[9] Perveen, S., Shahbaz, M., Guergachi, A., & Keshavjee, K. (2016). Performance analysis of data mining classification techniques to predict diabetes. *Procedia Computer Science*, *82*, 115-121.